Battling Antibiotic Resistance: Using Machine Prediction to Improve Prescribing*

Michael Ribers† Hannes Ullrich‡

October 1, 2018

PRELIMINARY - PLEASE DO NOT CIRCULATE OR CITE

Abstract

Machine learning methods are increasingly providing economists with opportunities to design welfare improving policies for problems that have prediction at their core. The alarming increase in antibiotic resistance due to improper use is one such opportunity. Human antibiotic intake is a main driver of antibiotic resistance and predicting bacterial presence early on is key to minimizing misuse. In this paper we evaluate how machine prediction can reduce wasteful overprescribing without diminishing health outcomes. Specifically, we train a machine learning algorithm on administrative data from Denmark to predict bacterial causes for urinary tract infections. The benchmark against which machine prediction must be evaluated is whether it can improve upon human decision making. Contrasting existing machine learning papers tackling prediction centered policy problems, in our setting patient test outcomes are observed independent of physician prescription choices. This allows us to directly evaluate health outcomes of prescription redistribution rules based on machine prediction. We find that redistribution rules based on a combination of machine prediction and physician autonomy can lower the existing level of antibiotic use by 10 percent without reducing the number of correctly treated bacterial infections. As Denmark is one of the most conservative countries in terms of antibiotic prescribing, this result is likely a lower bound of what can be achieved elsewhere.

*We benefited from very helpful feedback by and discussions with Rolf Magnus Arpi, Lars Bjerrum, Gloria Cristina Cordoba Currea, Greg Crawford, Tomaso Duso, Günter Hitsch, Ulrich Kaiser, Jenny Dahl Knudsen, Sidsel Kyst, Jeanine Miklós-Thal, Maria Polyakova, Stephen Ryan, Karl Schmedders, André Veiga, participants at the Annual Health Econometrics Workshop 2018 at Johns Hopkins University, and seminar participants at DIW Berlin and ESMT Berlin.

†University of Zurich - michael.ribers@business.uzh.ch

‡DIW Berlin, University of Zurich, Berlin Centre for Consumer Policies (BCCP), and CESifo - hannes.ullrich@business.uzh.ch.

1 Introduction

Empirical research in economics mostly focuses on uncovering causal treatment effects in order to give policy recommendations. However, economists are realizing more and more that many opportunities for welfare improving policies have prediction at the center, as stressed by Kleinberg et al. (2015) and Athey (2018). The alarming increase in antibiotic resistance worldwide constitutes one such policy problem.[1] It is a well established fact that human consumption of antibiotics is among the main drivers of antibiotic resistance. Correct prediction of bacterial causes to infection at the beginning of treatment, when the lack of diagnostic information is most pronounced, is a key challenge to minimize the misuse of antibiotics. When a patient first enters a doctor's office with symptoms of an infection, the quality of diagnosis and treatment depends among others on physician expertise, the availability and quality of diagnostic procedures, and information about the patient's medical history. A lack along any of these dimensions can lead to biased predictions of infection causes and ineffective treatment, with a substantial cost to society in the case of antibiotics.

In this paper, we evaluate how machine prediction can improve expert, general practitioner prescription decisions with the goal of reducing wasteful prescribing.[2] Specifically, we train a machine learning algorithm, a random forest, on rich, high-dimensional administrative data from Denmark to predict bacterial causes of urinary tract infections (UTI), one of the most common types of infections.[3] Bacterial UTI can be diagnosed with relative ease, by simple collection of urine samples and microbiological analysis. Yet, laboratory testing has the important and general practical limitation

[1] Antibiotics are medicines that treat bacterial infections by killing or otherwise inhibiting growth of bacteria in the body. Before the mass production of antibiotics following World War II, bacterial caused diseases now considered straightforward to treat had no medical remedy and were responsible for millions of lives lost. Antibiotics are losing their effectiveness due to antibiotic resistance, threatening to again render simple infections such as pneumonia or infections in wounds a fatal risk. The World Health Organization considers antibiotic resistance one of the greatest threats to global health given its immense costs in terms of life and treatment expenses (WHO 2012, 2014). Antibiotic resistance evolves when randomly mutating bacteria are exposed to evolutionary pressure generated by antibiotics in their local environment. Hence, governments and health organizations have broadly declared minimizing overuse of antibiotic treatment in humans a high priority path of action.

[2] In Denmark, general practitioners are responsible for the bulk of human antibiotic consumption, roughly 75 percent (Danish Ministry of Health 2017). Thus, we focus on general practitioners' prescriptions of systemic antibiotics, which are particularly important as they expose bacteria throughout the body.

[3] Foxman (2002) reports that almost half of all women contract a UTI once in their lifetime. In the United States, yearly UTI-related healthcare costs including workplace absences are estimated at $3.5 billion (Flores-Mireles et al. 2015). According to Bjerrum and Lindæk (2015), per year 10 percent of women receive antibiotic treatment for UTI.

that results arrive with a delay of several days, corresponding to nearly a complete course of antibiotic treatment. Until then, physicians must rely on their own risk evaluation.[4] During this waiting time, machine prediction can deliver a systematic bacterial risk assessment so far unavailable to the physician.

We show that machine predicted bacterial risk is highly correlated with realizations of bacterial UTI in out of sample patient test results. Yet, the ability to predict bacterial presence does not necessarily hold value in itself. The relevant criterion on which our tool needs to be evaluated is whether or not it can be used to improve human decision making. For this purpose, we model prescription decisions as a trade-off between the social cost of prescribing, which promotes resistance, and the curative effect to the patient in the event the patient symptoms have a bacterial cause. The model provides a framework to evaluate reassignment of antibiotic treatment based on a combination of the algorithm's prediction of risk and physician autonomy. We find that prescribing can be lowered by up to 10 percent with no reduction in the number of treated patients suffering from a bacterial infection. Careful evaluation of the machine learning assisted prescription rule leads us to conclude that we have indeed identified a potential to reduce prescribing by redistributing prescriptions from low risk to high risk patients. In addition to considering a policy that reduces overprescribing, we evaluate two alternative ways to optimize antibiotic prescribing. One alternative redistribution rule holds overall prescribing constant while aiming to increase proper anti-bacterial treatment. We show that such a rule has the potential to improve initial compliance between prescriptions and patients actually suffering from bacterial infections by 7.8 percent. Finally, we highlight the importance of combining machine learning with physician expertise by evaluating a redistribution rule based solely on machine predictions. We show that any rule that fails to include physician autonomy cannot lead to welfare increasing outcomes in the present context.

We use Danish data for three primary reasons. First, Danish administrative data are unique in the world in scope and interconnectivity. They cover a vast array of information including patients' and patient household members' personal background information such as detailed employment histories, as well as medical prescriptions and claims records, all of which are essential to conduct our analysis. The coherent use of unique person identifiers enables us to merge these data to individual laboratory test results. Second, due to these rich data we can demonstrate that similar results are achievable using a subset of predictors typically available to insurers or government agencies in developed countries. We also show that if that subset is too small, the achievable reduction in

[4]Throughout, when we write "physician", we mean general practitioners in an outpatient setting.

prescriptions is lower. Finally, Denmark is a country with an excellent record of low antibiotic use (Goossens et al. 2005) and with proper stewardship programs in place. Hence, the improvements we show for Denmark can be considered a lower bound on the attainable improvement elsewhere.

As all patients in our sample are tested at the initial consultation with a general practitioner, a crucial point is that we observe bacteria outcomes regardless of the physician prescription decision at that consultation. This allows us to avoid the selective labels problem confronting many prediction problems using observational data. When such data are generated by humans, they are the outcome of individuals' optimization problems and not a random data generating process. For example, in Kleinberg et al. (2017), where machine learning is used to improve judges' bail decisions, the selective labels problem occurs because the authors only observe crimes committed by released defendants. Predicting crime rates for the jailed is problematic as judges might have selected these individuals based on unobservables, thus creating biased machine predicted outcomes based on observables. We would face an analogous problem if we only observed test results for patients that were not prescribed antibiotics by the physician. Then, applying predictions based on these data to the prescribed without considering the selective labels problem would ignore physician selection, thus potentially invalidating our results. Hence, the unique opportunity to observe outcomes unconditional of physician prescription decisions is a key advantage of our analysis. This comes at the cost of only being able to use the patient population for whom physicians ordered laboratory test results at the initial consultation. Yet, in the context of antibiotic prescribing, our analysis holds empirical relevance and provides a health policy tool that effectively reduces overprescribing.

The remainder of the paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the institutional background and the data. Section 4 shows the results of the prediction algorithm relative to physician choices. Section 5 presents the framework for the design and evaluation of machine prediction-based antibiotic stewardship rules. Section 6 presents the achievable improvement in prescribing for a rule that aims to reduce overprescribing. Section 7 extends and discusses the results and Section 8 concludes.

2 Literature

This paper contributes to several strands of the literature. A growing set of empirical studies considers prediction policy problems (Kleinberg et al. 2015). Kleinberg et al. (2017) analyze the problem of predicting the risk of defendants' committing a crime in the context of judges' bail decisions.

They stress the importance of considering the role of unobservables, selective labels, and omitted payoffs to fairly compare machine prediction to human decisions. We largely rely on their proposed approach but our data allow us to shed further light on the role of unobservables and to consider a richer set of policy improvements. Chalfin et al. (2016) predict worker productivity to improve police hiring practices and teacher tenure decisions. In these contexts, they stress the importance of considering decision makers' potentially complex payoff functions to draw policy conclusions based on machine prediction. Misspecifying payoff functions, for example omitting important payoff dimensions, will lead to biased and potentially harmful policy outcomes. While medical decision making typically has non-trivial payoff functions, the decision with regard to antibiotic prescribing is fairly straightforward as physicians trade off the private benefit of curing a sick patient and the social cost of promoting antibiotic resistance. We tackle the main challenge in this decision context, heterogeneity in patients' sickness disutility, by investigating the observable characteristics of individuals affected by our prescription redistribution rules and by adapting these rules accordingly.

In medicine, machine learning for prediction has received much attention driven by hopes that electronic health records, insurance claims, public registries, as well as genomics databases will help inform medical decision making. Obermeyer and Emanuel (2016) describe how machine learning will improve prognosis in clinical practice in the near future. For example, machine learning methods have been used and are expected to be used increasingly to predict organ failure or mortality, to automatize the interpretation of medical imaging such as mammograms, to automatize 24-hour monitoring in critical care, as well as to improve diagnostics. Chen and Asch (2017) highlight the challenges arising from the many complicating factors in the medical context. For example, predicting cancer treatment success in the distant future is a very difficult task based on historical health data. Google Flu Trends received much attention but had very little success when they attempted to predict the prevalence of influenza by combining large-scale online search data with a small sample of influenza cases. One core lesson from this low performance is that collecting and combining relevant types of data is crucial (Lazer et al. 2014). Currie and MacLeod (2017) consider expert decision making in the context of physicians' procedural choice of Cesarean sections in child birth. They find that administrative data allow to identify physicians who make poor decisions leading to negative health outcomes. Likewise, in antibiotic treatment decisions, expertise heterogeneity when diagnosing bacterial infections is expected to be a main driver of antibiotic misuse. We propose implementing machine prediction for first-contact diagnostics in a setting in which rapid medical diagnostics are not available but a large set of relevant predictors of test

outcomes are.[5] In addition to being able to predict bacterial infections accurately, we consider fair comparisons between physicians' decisions and decisions based on our machine predictions.

A large economic and public health literature considers antibiotic resistance as an externality of antibiotic consumption (Brown and Layton 1996, Laxminarayan and Brown 2001, Laxminayaran and Weitzman 2002, Rudholm 2002, Elbasha 2003, and Herrmann and Gaudet 2009). The list of proposed mechanisms that policy can leverage is extensive, including reforming patent systems to incentivize new drug development (Eswaran and Gallini 2018) and demand-side measures such as prescription surveillance and stewardship (Laxminarayan et al. 2013), general practitioner competition (Albert 2015, Bennett et al. 2015), financial incentives for physicians (Yip et al. 2010, Currie et al. 2014, Das et al. 2016), education programs (Arnold and Straus 2005, Butler et al. 2012), peer effects (Kwon and Jun 2015), social norm feedback (Hallsworth et al. 2016), and reducing diagnostic uncertainty by investing in rapid diagnostics (Laxminarayan et al. 2013). We contribute to this literature by providing a tool that leverages machine learning methods and large-scale data to reduce both over- and underprescribing when treatment choices are made under uncertainty about the cause of infections.

This paper does not consider many of the details involved in the antibiotic treatment decisions of (urinary tract) infections that medical experts must consider (Hooton 2012, Grigoryan et al. 2014, Bjerrum and Lindbæk 2015). We focus on a physician's simple, focal choice situation when initially diagnosing a patient with UTI symptoms: predicting whether or not an infection has a bacterial cause, then making an appropriate treatment decision. Randomized interventions, such as in Hallsworth et al. (2016), are promising as they significantly decreased the number of antibiotic prescriptions. Yet, to our knowledge, there is a lack of evidence that considers the effect of such interventions on the quality of treatment decisions. We directly evaluate and optimize the use of antibiotics, leading to a reduction in overall use while holding the number of correct treatments constant.

[5]These predictors are retrieved from patients' medical claims histories, past microbiological test results, personal characteristics, as well as patients' household members' background and medical data.

3 Data

3.1 Institutional Setting

We begin the data section with an overview of the institutional setting in Denmark where we focus on important regulation that impacts general practitioner decision making. The institutional background is important to understand the scope of our results outside the Danish context. Denmark has a universal and tax financed single payer health care system. Systemic antibiotics are classified as prescription drugs with general practitioners as the primary gatekeepers. General practitioners are responsible for prescribing approximately 75 percent of the human consumed systemic antibiotics in Denmark (Danish Ministry of Health 2017). Although general practitioners operate as privately owned businesses, all fees for services are collectively negotiated between the national union of general practitioners and the public health insurer. Importantly, physicians do not generate earnings by prescribing drugs to patients who have to purchase their prescriptions at local pharmacies. Pharmacies earn a fixed fee per prescription processed regardless of the prescription drug price or other drug attributes, for example branded versus generic drugs. Prescription drugs are subsidized but patients co-pay a fraction of the list price depending on their cumulative yearly prescription drug expenditures. The Danish market for prescription drugs is highly regulated resulting in uniform pricing at pharmacies nationwide. In general, antibiotic treatment is cheap.[6] Lastly, patients are allocated to general practitioners by a list-system within a fixed geographic radius around their home address. Patients can switch physicians from their initial assignment at a small cost but most stick with their assigned general practitioner.

3.2 Danish National Registries

We use administrative data provided by Statistics Denmark covering all citizens in Denmark between January 1st, 2002 and December 31st, 2012. For each individual, we observe the complete prescription history of systemic antibiotics (*Lægemiddeldatabasen*), hospitalizations (*Landspatientregisteret*), general practitioner medical claims (*Sygesikringsregisteret*), and a comprehensive set of socioeconomic and demographic variables linkable to both patients and physicians via unique patient and physician identifiers. The prescription data contain information on all systemic antibiotics prescribed and purchased in Denmark. In total, the data contain 35.3 million systemic

[6] Antibiotic treatment typically lasts three to seven days and costs do not exceed 30 DKK ($5) per defined daily dose for the most commonly used antibiotics against UTI.

antibiotic prescriptions and among others include variables for the date of purchase, price, size of subsidy, anatomical therapeutic chemical classification (ATC), drug name, indication, and defined daily dose (DDD). Given Denmark's universal and tax financed single payer health care system, the insurance claims data covering general practitioner clinics include all physician services provided to the patient. The claims data are comprised of approximately 100 million observations per year and include variables identifying service type, physician fee, and consultation timing. The claims data allow us to identify pregnant women from mandatory pregnancy-associated examinations. This is important as pregnant women are considered complicated UTI cases, a point we discuss in detail following our main results. A further important service contained in the claims data are urine nitrate sticks tests which provide physicians with an indication, albeit imperfect, of whether UTI symptoms have a bacterial cause. Finally, the national hospitalization data comprise all patient contacts with hospitals, including ambulatory visits. The data contain observations on hospitalizations of approximately 3.5 million unique individuals per year in our 10 year observation period and include information on hospitalization date, diagnosis, hospitalization type, and bed days.

3.3 Clinical Microbiological Test Results

We have acquired test results for the time period between January 1st, 2010, and December 31st, 2012, from clinical microbiological laboratories at Herlev hospital and Hvidovre hospital, two major hospitals in Denmark's capital region covering a catchment area of roughly 1.7 million people. The laboratory test data are central because they reveal bacterial presence in a urine test sample, the outcome we aim to predict. Overall, the data contain 2,579,617 biological samples submitted for testing in the capital region by general practitioner clinics and hospitals. Urine samples constitute 477,609 samples out of which 153,323 are submitted by general practitioners.[7] Using patient identifiers, we link these test results to the administrative data described above.

For the purpose of this paper, we assume that whenever a urine sample was submitted for laboratory analysis, the corresponding patient was reporting UTI related symptoms. Approximately one out of three urine samples contain bacteria isolates, both overall and among the general practitioner submitted samples. In addition to a patient and general practitioner clinic identifier, the data contain a test type, test date, sample arrival date, test result response date, isolated bacteria, and a list of antibiotic molecule-specific resistances if bacteria were isolated. The typical test procedure

[7]Urine samples from general practitioners are identified by department IDs UP1, UP2, UP3 and U-2, where the latter is crossed with general practitioner clinic codes as it contains both hospital and general practitioner samples.

takes two to four days during which general practitioners are uninformed about the test result. Observing the precise timing of test acquisitions, prescription purchases, and test responses allows us to define and evaluate physicians' treatment decisions during the waiting period, i.e. before the physician was informed about the test result.

3.4 Data Partitions to Train and Evaluate the Machine Learning Algorithm

To create our full data set, we combine test outcomes, Y, with the administrative datasets containing extensive patient-specific explanatory variables, X, that serve as input in our prediction algorithm. Altogether, X comprises $1,215$ variables which are in principle all observable to the physician at the time of the consultation. For exposition, we collect our variables in groups, where for each patient we distinguish between (i) the consultation date and patient demographics, (ii) patients' past prescriptions and the past three months' average consumption in their home municipalities, (iii) patients' past microbiological test results, and (iv) patients' past hospitalizations. Further, the Danish registry data include information about patients' household members categorized by family relation. This allows us to include the additional explanatory variables (v) household members' past prescriptions, (vi) household members' past microbiological test results, and (vii) household members' past hospitalizations by family status relative to the patient in question.[8]

We restrict the outcome to test observations submitted by general practitioners and keep only patients who did not receive any systemic antibiotic prescriptions in the 28 days prior to the respective test date.[9] We make this restriction to focus on such consultations that constitute a first contact with a physician within a patient's treatment spell. In these situations, physicians do not hold current test result information and must prescribe under uncertainty. In addition, by considering only initial consultations, we exclude potentially complicated treatment spells where patients are tested in later stages. We also avoid patients in longterm treatment, potentially due to severe antibiotic resistance problems.

As we aim to accurately predict out of sample, we separate the data into a 'training' and a 'hold out' partition. After fitting the prediction function using the training data, we evaluate its out of sample predictions on the hold out partition. The separation of the data is performed on the municipality level such that for municipalities belonging to the hold out data all physicians

[8]The definition of a family unit includes definition of parents and singles or couples with or without kids. See the definition of familie_id at www.dst.dk for further details.

[9]We do not exclude historical non-general practitioner information from the explanatory variables as these may be important predictors for bacterial cause of infections.

and their complete patient pools are excluded from the training data. Figure 1 marks the training municipalities in light gray and the hold out municipalities in black. Hence, we have ensured that the algorithm faces completely isolated geographic regions in the hold out data, a similar exercise as exporting the results to other parts of Denmark not covered by the two hospital laboratories, or beyond.

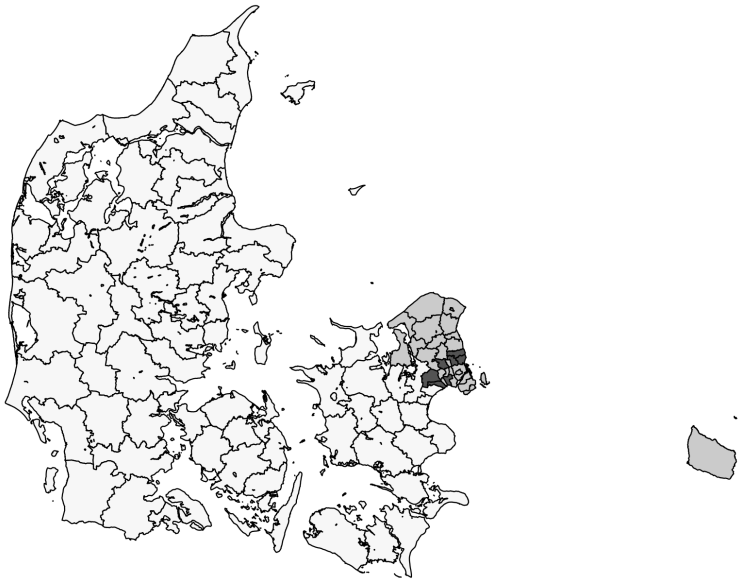


Figure 1: Overview of municipalities in Denmark covered by the clinical microbiological laboratories in our data conditional on our partition of the training municipalities (light gray) and the hold out municipalities (black).

Table 1 reports the level of a few key demographic variables: income, age, gender, civil status, immigration status, education, and the number of physicians across the training and hold out partitions. Differences among the tested are reasonably small with the hold out partition showing slightly more married, older, and less educated patients, which is to be expected as Copenhagen is not included in the hold out municipalities. An overview of the prevalence of the isolated genera in the training and the hold out partition is shown in Table 2 in Appendix A. Escherichia is the most prevalent bacterial genus in the joint data appearing in 64.2 percent of the cases where bacteria was isolated. Enterococcus (9.2 percent) is the second most frequent appearing genus among the isolates, followed by Staphyloccocus (7.7 percent), Streptococcus (5.7 percent), Klebsiella (5.5 percent) and further low frequency genera. It is important to note that the training and hold out partitions show roughly similar rates of genera and there is thus no indication of any structural differences between the partitions.

Table 1 Observables for the Training Data and the Hold Out Data

Observables	Training Data	Hold out Data
Yearly income, DKK	237,312	240,918
Age, years	42.3	49.9
Female	0.84	0.79
Married	0.34	0.44
Immigrant	0.17	0.15
Tertiary Education	0.35	0.26
# Physicians	402	120
Observations	90,946	17,112

4 Machine Prediction

Following standard machine learning practice, we train an algorithm relating patient observables, X, to the bacterial outcomes, Y, where the latter indicates if the patient was tested positive for *bacterial* presence in a urine test. Importantly, X only contains historical observations relative to the patient observation test date. We include all tests submitted at initial visits in a patient spell in which the patient reports UTI related symptoms, as indicated by the need of bacterial screening. We choose a random forest algorithm (Breiman 2001) over alternatives such as logistic regression, LASSO, or neural nets due to its parsimony, low computational cost, and ability to uncover highly flexible, nonlinear functions in a high-dimensional feature space while avoiding overfitting. The ability to accurately predict bacterial presence does not necessarily hold value in itself. The benchmark against which our tool needs to be evaluated is whether or not it can be used to improve physician decision making. Hence, we begin by inspecting the quality of machine predictions and the disagreements with physicians' prescription decisions. In Section 5, we then consider physician choices and potential for improvements in further detail.

4.1 Machine Learning Performance

The random forest is an ensemble of recursive binary decision trees applied to bootstrapped versions of the training data. A decision tree represents a partition of the data as a sequence of binary splits over individual variables in X so that each split is determined by the homogeneity of the predicted

variable Y in the created partitions. When all splits have been made, a simple model, in our case a constant 0 or 1, is universally fitted to the remaining observations in each final partition, or leaf, of the tree. The random forest predictions, $m(X)$, is determined as the democratic vote among trees.

A variety of final binary predictions can now be accomplished by comparing the majority vote to a classification threshold. In common machine learning applications this threshold is simply set to pick the majority, i.e. set to 0.5, but for physicians evaluating bacterial presence different thresholds might constitute sufficient risk. Any choice of classification threshold leads to different predicted outcomes that, combined with the true outcome, can be sorted into four categories: correctly predicted positives, correctly predicted negatives, falsely predicted positives, and falsely predicted negatives. As the classification threshold varies over $[0, 1]$, the receiver operating characteristic curve is a plot of the algorithm's correctly predicted positives out of all actual positives against falsely predicted positives over all actual negatives. Hence, it is a measure of the trade-off between getting positives right and negatives wrong. Figure 13 in Appendix B shows the receiver operating curve for our application. Generally, the closer the receiver operating curve is to the (0,1) corner, the better prediction quality. A receiver operating curve along the 45-degree diagonal represents random guessing. Extending this insight leads to a common metric by which to measure prediction accuracy: the area under the receiver operating characteristic curve, the AUC. The interpretation of the AUC is that 0.5 represents no better than random guessing conditional on the sample means, a value below 0.5 would be worse than random guessing, while a value of 1 represents perfect prediction. Our bacterial prediction function has an AUC equal to 0.726. Kleinberg et al. (2017) report a comparable AUC of 0.707 in the crime risk context. We report further metrics for the accuracy and evaluation of the random forest algorithm in Appendix B.

4.2 Predicting Bacterial Presence

The left panel of Figure 2 plots the average actual bacterial test results in the hold out partition against the average algorithm predicted risk. Every circle represents a bin containing 100 patients where patients are assigned to bins by their predicted risk percentiles. Outcomes are centered around the 45 degree line throughout the risk distribution which shows that the algorithm on average correctly predicts bacterial risk. Out of the predicted riskiest 100 patients in the hold out data, 87.6 percent are tested positive for bacteria following the initial consultation with the physician. Equivalently, the observed bacterial UTI rate for the 100 least riskiest patients is 3.0 percent. The right panel of Figure 2 plots the physicians' prescription rate prior to obtaining a test

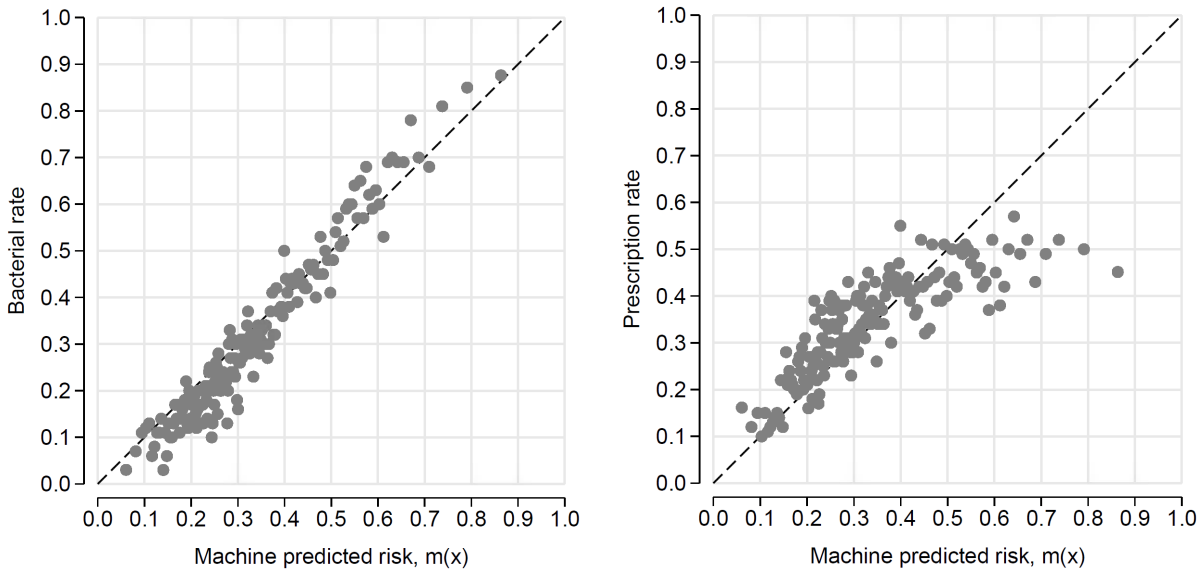


Figure 2: Actual bacterial test results related to machine prediction of bacteria (left) and physician prescribing prior to obtaining the patient test result related to machine prediction of bacteria (right) both evaluated on the hold out partition. Each circle represent averages over bins containing 100 patients sorted by predicted risk.

result against the algorithm's predicted risk. Again, circles represent averages over bins containing 100 patients sorted by the algorithm's predicted risk. Physicians seem to evaluate low risk patients correctly on average as the prescription rate and predicted risk appear well correlated in the low risk range. However, in the range of high predicted risk, the physicians' prescription rate flattens out with most observations remaining within the 40 to 60 percent range. Hence, the physicians and the algorithm seem to disagree on the high risk patients where in particular it appears that physicians have difficulty identifying high risk patients. The fact that the algorithm performs well throughout the risk range is a sign that improvements to physician decision making might be attainable with the algorithm's help.

5 Evaluating Machine Prediction Based Policies

In this section, we construct a framework that enables us to design and evaluate the impact of antibiotic prescription policies based on machine predictions. Inspired by Kleinberg et al. (2017), we consider a policy maker's problem as solving a trade-off between patient sickness and the social cost of prescribing, that is, promoting antibiotic resistance. We turn our attention to prescription

decisions at a patient's first visit of a potentially longer treatment spell and propose counterfactual prescription rules based on machine predictions. These rules delay prescribing for some patients until test results are received and give prescriptions instantly to others, hence overruling observed physician choices for a subset of patients.

A common challenge when evaluating counterfactuals in machine learning applications is the selective labels problem, similar to sample selection causing bias in causal models.[10] By focussing on the sample of prescription occasions that all include microbiological testing, test outcomes are observed for all patients regardless of the physician's prescription decision. This means we do not face the selective labels problem and can evaluate antibiotic stewardship rules directly, allowing us to construct and evaluate a large range of redistribution rules based on machine prediction. The disadvantage to our approach is that we cannot claim the generalizability of our results to prescription occasions that did not include patient microbiological testing. However, the tested cases are significant in number, accounting for approximately 15 percent of all initial UTI consultations. Further, when a physician decides to test, the value of diagnostic information is presumably high so that the machine prediction-based policies proposed here improve upon situations in which physicians are making decisions under significant uncertainty.

5.1 Definition of Policies

Before we introduce the framework we use to evaluate the benefits of policies based on machine predicted risk, for tractability we limit the set of policies considered for evaluation. First, we only consider policies that do not discriminate between individual physicians. Second, any policy must *either* enforce prescribing, enforce no prescribing, *or* leave prescribing to physician autonomy for all patients with similar realizations of machine predicted risk. Physician autonomy over a range of machine predicted risk means that a proposed policy does not affect or change physician choice for the set of patients in this range. With these two general limitations in place, we turn our attention to the machine predictions on the hold out data in order to further refine polices to be evaluated.

[10] In the bail decision context in Kleinberg et al. (2017), the selective labels problem manifests itself in that crime outcomes are only observed for released defendants. If judges assess risk based on unobservables and make release decisions accordingly, crime risk for the jailed cannot be compared to crime risk of the released, even conditional on observables. Hence, any counterfactual release rule based on the released must be evaluated with care. One way around this issue is to evaluate a particular form of policy, namely release rules that jail similar defendants released by one judge but jailed by another. Two necessary assumptions for this approach are random assignment of defendants to judges and varying leniency in release decisions across judges.

In particular, we examine the machine predictions vs actual bacterial UTI realizations conditional on physician prescribing as shown in Figure 3.

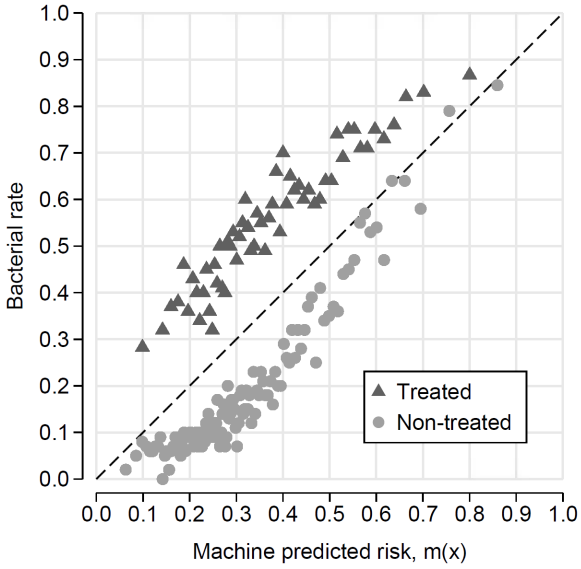


Figure 3: Machine predictions vs actual bacterial UTI realizations conditional on physician prescribing. Markers represent averages over bins of 100 patients sorted by predicted risk.

Two main insights follow from Figure 3. First, conditional on a level of machine-predicted risk, physicians are on average able to prescribe to patients with more frequent realizations of bacterial UTI. Two reasons for this could be physician expertise and diagnostic unobservables, the latter among other covering in-house diagnostic tests such as nitrite dipsticks or microscopy.[11] This finding suggests that it is optimal to include physician autonomy for some range of machine predicted risk. Second, even though physicians are able to better identify patients with bacterial UTI conditional on the level of machine-predicted risk, this does not imply that physician decision making cannot be improved in some ranges of machine predicted risk. To see this, note that in Figure Figure 3

[11]Nitrite dipstick can detect bacteria that transform Nitrate to Nitrite. In the hold out data, the detectable genera are Escherichia, Enterobacter, Klebsiella, Citrobacter, and Proteus. The non-detectable genera are Staphylococcus, Pseudomonas, Enterococci, Acinetobacter, and Streptococcus. Inspecting prescription choices separately by dipstick-detectable and non-detectable bacterial species isolated in laboratory tests allows to investigate whether physicians select on nitrite dipstick test results. While patients with dipstick-detectable bacteria have a higher prescription rate, 64 percent, than patients with non-dipstick-detectable bacteria, 55 percent, the difference is moderate. This suggests that dipstick test results leave significant uncertainty, which is consistent with evidence reported in the medical literature (Devillé et al. 2004).

the 100 patients with the lowest predicted risk among the patients that received a prescription in the hold out data (the left most triangle) on average had a bacterial caused UTI in 28 percent of the cases. That is, a total of 100 patients received prescriptions in this group in order to only treat 28 patients with bacterial caused UTIs. We define overprescribing as any prescription to a patient not suffering from bacterial UTI and observe that overprescribing on average decreases among the treated as machine predicted risk increases. Similarly, the 100 riskiest patients not receiving prescriptions (the right most circle) have an average actual bacterial prevalence of 84 percent. We define underprescribing as any patient suffering from bacterial UTI that did not receive antibiotics and note that underprescribing on average decreases as machine predicted risk decreases among the non-treated. In conclusion, overprescribing is predominant at lowest machine predicted risk becoming less prevalent as machine predicted risk increases; and underprescribing is predominant at highest machine predicted risk becoming less prevalent as machine predicted risk decreases. This gives rise to a specific form of machine prediction-based policies, ρ^S, of the following form:

$$\rho^S(x_i; k_L, k_H) = \begin{cases} 0 & \text{if } m(x_i) < k_L, \\ \rho^j & \text{if } k_L \leq m(x_i) \leq k_H, \\ 1 & \text{if } k_H < m(x_i), \end{cases} \tag{1}$$

where x_i is patient i's explanatory variables, $m(x_i)$ is machine predicted risk, and k_L and k_H are policy thresholds to be determined based on machine predicted risk for the hold out data and the policy maker's objective, which we turn to in the next section.

5.2 Policy Maker Utility

We assume that a policy maker is aware of the resistance-promoting effect of antibiotic consumption such that she considers a fixed social cost for any antibiotic prescription. The policy maker also cares about patient health. Prescribing antibiotics has a curative effect if and only if the patient suffers from a bacterial infection.[12] Hence, in order to optimize the machine assisted policies given by equation (1), we assume a policy maker weighs the benefits to patient i from antibiotic treatment against the social cost of prescribing as follows:

$$\pi^S(p_i; y_i) = -a_S(1 - y_i p_i) - b_S p_i, \tag{2}$$

[12] Antibiotic treatment has no curative effect when the cause of the infection is viral, fungal or otherwise related to non-bacterial causes.

where $p_i \in \{0,1\}$ is the prescription choice to patient i, $y_i \in \{0,1\}$ determines whether patient i's infection has a bacterial cause, and a_S, b_S represents the policy maker's weight on patient sickness and the social cost of prescribing, respectively. We assume that $0 < b_S < a_S$ which restricts the policy maker's preferences such that prescribing is optimal when an infection is known to be bacterial with certainty and that prescribing is not optimal when an infection is known to not be caused by bacteria. In order to evaluate the policy maker's payoff gain (or loss) from implementing the policy given in equation (1), we compute the payoff gain (or loss) compared to full physician autonomy over all patient observations in the hold out partition:

$$\begin{aligned}\Pi(k_L, k_H) &= \sum_i \left(\pi_S(y_i, \rho^S(x_i; k_L, k_H)) - \pi_S(y_i, \rho_i^J)\right) \\ &= a_S \underbrace{\sum_i y_i(\rho^S(x_i; k_L, k_H) - \rho_i^J)}_{\Delta\text{treated bUTIs}} - b_S \underbrace{\sum_i \left(\rho^S(x_i; k_L, k_H) - \rho_i^J\right)}_{\Delta\text{antibiotic use}},\end{aligned} \tag{3}$$

where ρ_i^J is the observed physician prescription decision for patient i. In order to determine the optimal k_L and k_H, we define the following two objectives:

- **Minimize antibiotic use** holding overall bacterial UTI treatments constant.
- **Maximize bacterial UTI treatment** holding overall antibiotic use constant.

We choose the above two restricted objectives because they have the potential to ensure a positive payoff to the policy maker regardless of the chosen a_S and b_S. To see this, note that by holding bacterial UTI treatments constant the first term of equation (3) is zero and so, if we manage to reduce prescribing with machine predictions under this constraint, the second term will have a positive impact on utility as long as $b_j > 0$. Similarly, by holding antibiotic use constant the second term of equation (3) is zero. If we manage to increase bacterial UTI treatments under this constraint using machine predictions, the first term will have a positive impact on utility as long as $a_j > 0$.

For specific values of policy makers' preference parameters a_S and b_S it can be optimal to reduce antibiotic use further than under the objectives defined above. For example, under certain preference parameter values one may be willing to accept a reduction in the number of treated bacterial UTIs to reduce overall prescribing further or to accept larger numbers of prescriptions to treat more bacterial UTIs. The rate at which policy makers desire to do so depends on the relationship between a_S and b_S in combination with the rate at which machine predictions improve the trade-off between over- and underprescribing. We do not observe policy makers' preferences

which are, in general, challenging to identify and estimate. Instead, we identify the minimum positive welfare improvement achievable without knowledge of a_S and b_S.

6 Results

We present the results from implementing the policy rules in section 5.2 as the percentage change in antibiotic use in the hold out data set:

$$\Delta\rho(k_L, k_H) = \frac{\sum_i \left(\rho_i^J - \rho^S(x_i; k_L, k_H)\right)}{\sum_i \rho_i^J} \tag{4}$$

and the percentage change to bacterial UTI treatments:

$$\Delta\text{bUTI}(k_L, k_H) = \frac{\sum_i y_i \left(\rho^S(x_i; k_L, k_H) - \rho_i^J\right)}{\sum_i y_i \rho_i^J}, \tag{5}$$

the percentage equivalents to the difference terms in equation (3).

Minimize antibiotic use

The optimal thresholds (k_L^*, k_H^*) for the objective to minimize antibiotic use are found by maximizing equation (4) under the constraint that equation (5) equals zero. Figure 4 shows the potential gain from redistribution of prescriptions from patients with a predicted risk below k_L^* to patients with a predicted risk above k_H^*. The left panel of Figure 4 shows (k_L^*, k_H^*) and the set of (k_L, k_H) that sets equation (5) to zero; and the right panel shows the attainable percentage reduction in prescribing for all possible $k_L \in [0, 0.34]$ along this path. The maximum reduction in antibiotic prescribing is 10.0 percent of the observed prescription level which is achieved by a redistribution rule that sets $k_L^* = 0.279$ and $k_H^* = 0.531$. Having determined k_L^* and k_H^*, we bootstrap the hold out data to estimate a sampling error of the estimated reduction in antibiotic use. For 100 bootstrapped hold out samples, the standard error on the percentage reduction in antibiotic use is 0.82 and the 95% confidence interval on the normal distribution is $[8.39, 11.6]$, suggesting sampling error is small. Although the number of patients treated for bacterial UTI is constrained to be constant, this constraint does not hold precisely. The bootstrapped 95% confidence interval for deviations from zero change in treated bacterial UTI is [-2.2, 2.2] percentage points.

Figure 5 shows the patient distribution as a function of predicted risk conditional on treatment. Patients affected by redistribution are the patients to the left of k_L^* and right of k_H^*; and prescriptions are removed on the left panel and added on the right. The redistribution rule balances the number

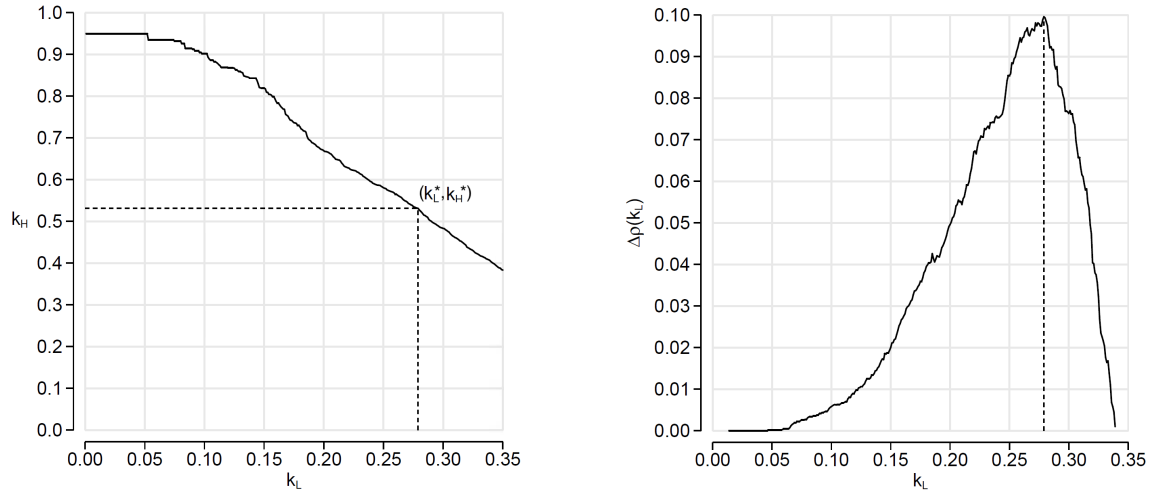
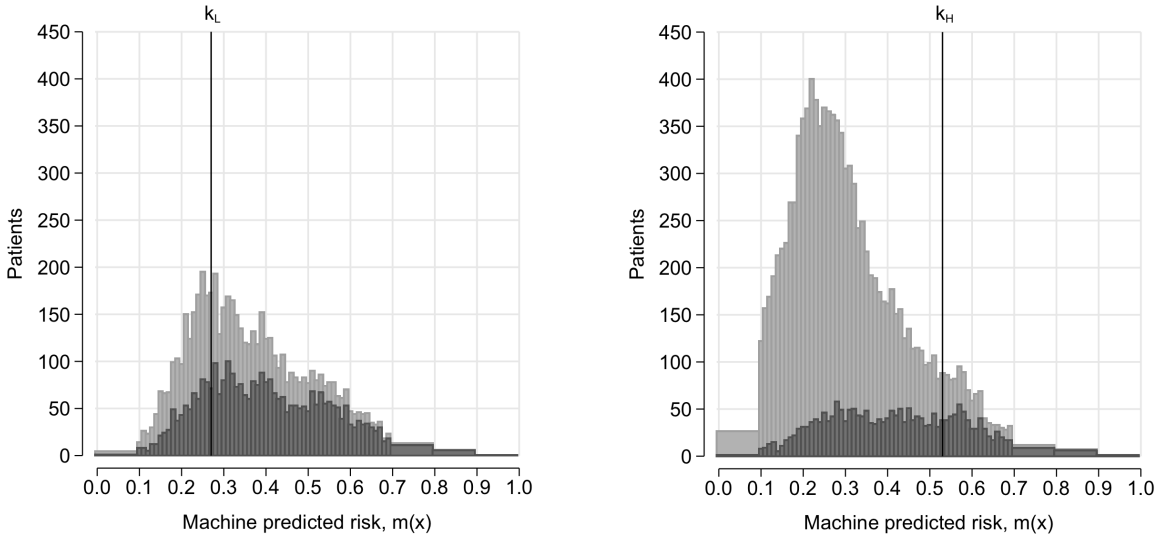


Figure 4: Levels of k_L and $k_H(k_L)$ holding bacterial UTI treatments constant and optimal (k_L^*, k_H^*) that minimize antibiotic use (left). Minimizing antibiotic use for fixed number of correctly treated bacterial UTIs as a function of k_L (right).



(a) Treated **(b)** Non-treated

Figure 5: Distribution of treated and non-treated patients by machine predicted risk with optimal (k_L^*, k_H^*) reducing antibiotic use with constant bacterial UTI treatments. Prescriptions delayed left of k_L^* and given to the right of k_H^*. Frequencies in the distribution tails are averaged and collapsed into 10th percentiles for anonymity reasons.

of bacterial UTI cases (dark grey) that have prescriptions shifted while reducing the non-bacterial cases (light grey) as much as possible. We can see that the redistribution rule gives prescriptions

to patients in the long tail of the predicted risk distribution and that approximately 30 percent of the prescribed have their treatment removed, a relatively large intervention.

Maximize bacterial UTI treatment

The optimal thresholds (k_L^{**}, k_H^{**}) for the objective to maximize bacterial UTI treatments are found by maximizing equation (5) under the constraint that equation (4) equals zero. Figure 6 shows the

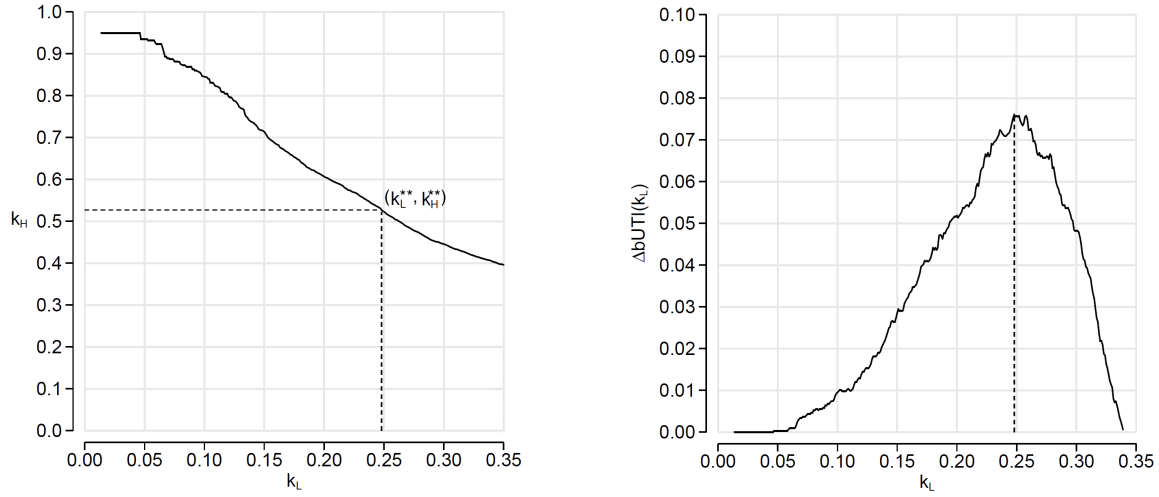


Figure 6: Levels of k_L and $k_H(k_L)$ holding antibiotic use constant and optimal (k_L^{**}, k_H^{**}) that maximize bacterial UTI treated (left). Increase in treated bacterial UTI holding antibiotic use constant as a function of k_L (right).

potential gain from redistribution of prescriptions from patients with a predicted risk below k_L^{**} to patients with a predicted risk above k_H^{**}. The left panel of Figure 6 shows (k_L^{**}, k_H^{**}) and the set of (k_L, k_H) that sets equation (4) to zero. The right panel shows the attainable percentage increase in the number of treated bacterial UTI patients for all possible $k_L \in [0, 0.34]$. The maximum increase in treated bacterial UTIs is 7.6 percent which is achieved by a redistribution rule that sets $k_L^{**} = 0.248$ and $k_H^{**} = 0.527$. For these optimal k_L^{**} and k_H^{**} and 100 bootstrapped hold out samples, the standard error on the percentage change of treated bacterial UTIs has a 95% confidence interval on the normal distribution of $[5.6, 9.6]$. Holding antibiotic use constant is only approximately achieved in the bootstrap hold out samples, with a 95% confidence interval of [-1.8, 1.7] percentage points. Figure 7 shows the patient distribution as a function of predicted risk conditional on treatment. Again, patients affected by redistribution are the patients to the left of k_L^{**} and right of k_H^{**}; and prescriptions are removed on the left panel and added on the right.

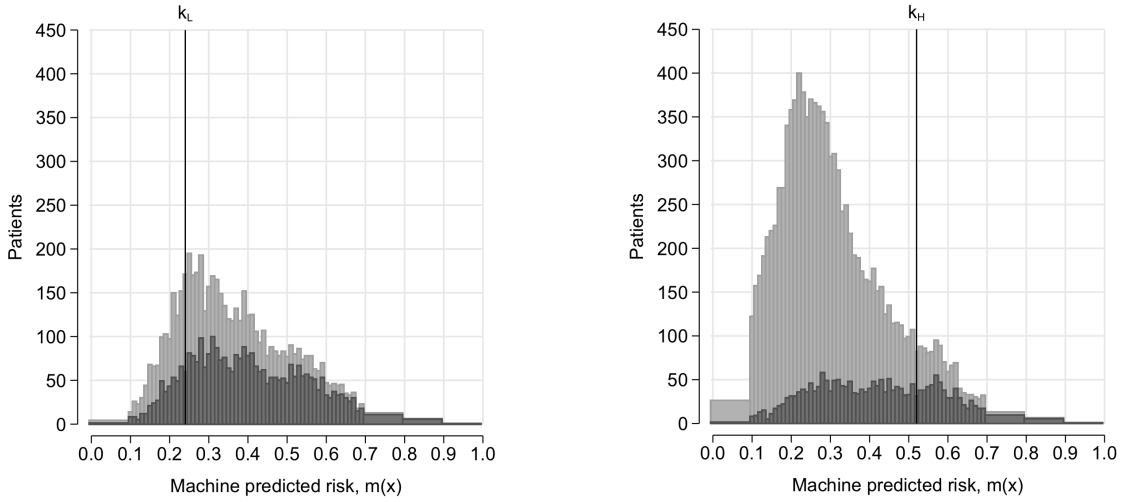


Figure 7: Distribution of treated and non-treated patients by machine predicted risk with optimal (k_L^{**}, k_H^{**}) maximizing bacterial UTI treatments holding antibiotic use constant. Prescriptions delayed left of k_L^{**} and given to the right of k_H^{**}. Frequencies in the distribution tails are averaged and collapsed into 10th percentiles for anonymity.

Are physicians needed?

A special case of the general prescription rule in equation (1) is when $k_L = k_H$, i.e. when prescription policy is based solely on machine predictions, completely ignoring physician information and decision making. Such a rule becomes a step function where prescriptions are never given below the cut-off $k_M \equiv k_L = k_H$ and always given above:

$$\rho_i^M = 1 \text{ if and only if } m(x_i) \geq k_M \tag{6}$$

We find that such a rule is inferior to versions of equation (1) that includes physician autonomy. In particular, such a rule fails to implement welfare improvements for general specifications of policy maker preferences. Specifically, the rule is not able to reduce antibiotic use without also reducing treated bacterial UTIs; or, equivalently, it is not able to increase treated bacterial UTIs without also increasing antibiotic use.

The left panel of Figure 8 shows the implementation of the prescription rule ρ^M with the optimal $k_M = 0.35$. All patients in the hold out data with predicted risk below 0.35 are restricted from receiving a prescription and all patients with a predicted risk above or equal to 0.35 are prescribed antibiotics. The right panel of Figure 8 shows the outcomes of ρ^M in terms of equations (4) and (5) for all $k_M \in [0, 1]$. It is seen that it is not possible with ρ^M to reduce antibiotic use, $\Delta\rho(k_M, k_M) > 0$,

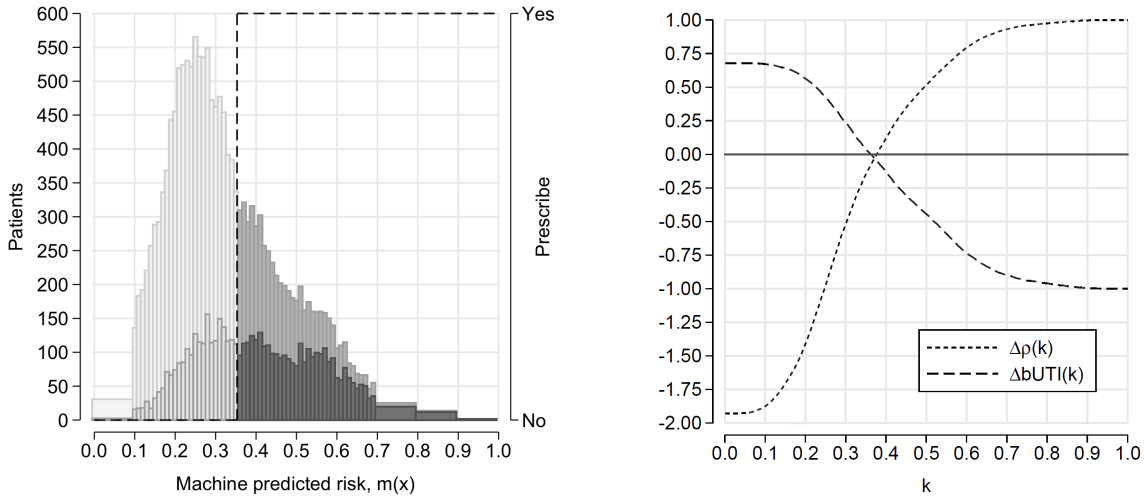


Figure 8: Machine redistribution policy without physician autonomy (step function) enforcing prescribing above 35 percent machine predicted risk. Frequencies in the tails of the distribution are averaged and collapsed into 10th percentiles for anonymity reasons.

without also decreasing the number of treated bacterial UTIs, $\Delta\text{bUTI}(k_M, k_M) < 0$. Equivalently, it is not possible to increase the number of treated bacterial UTIs, $\Delta\text{bUTI}(k_M, k_M) > 0$, without increasing antibiotic use, $\Delta\rho(k_M, k_M) < 0$. Therefore, we conclude that such a rule cannot ensure welfare improvement for general policy maker preferences as in the proposed rules ρ^S including physician autonomy. This suggests that in this setting of antibiotic prescribing, even with large amounts of detailed, individual-specific data, machine predictions only lead to general and significant welfare improvements when combined with physician expertise.

7 Discussion

We have shown that machine prediction-based prescription policies can deliver substantial welfare gains. In this section we will discuss the main findings by exploring potential misspecification of physician preferences as an alternative explanation for discrepancies between the physicians' prescription choices and the algorithm's decision rule. We further investigate the data requirements required for these policies to assess their generalizability to other countries and contexts where data availability may be more limited.

7.1 Physician Preferences: Patient Heterogeneity in Sickness Cost

Until now we have ignored one potentially problematic assumption in our model: sickness cost, a_S, is constant across all patients in the policy maker's utility specification. Some patients might be high cost, not to be confused with high risk, in the sense that the health consequence of having a bacterial UTI would be much more severe for such patients compared to others. One important example are pregnant women. For example, Schieve et al. (1994) show that UTIs during pregnancy are positively associated with adverse outcomes such as low birthweight, prematurity, maternal anemia, and others. Foxman (2002) in a review reports that a UTI during pregnancy leads to elevated risk of kidney infection, fetal mortality, and premature delivery. Schwandt (2018) finds that in-utero exposure to influenza infections leads to premature delivery and low birthweight in the short-run and a nine percent reduction in earnings and a 35 percent increase in welfare dependence in the long-run.

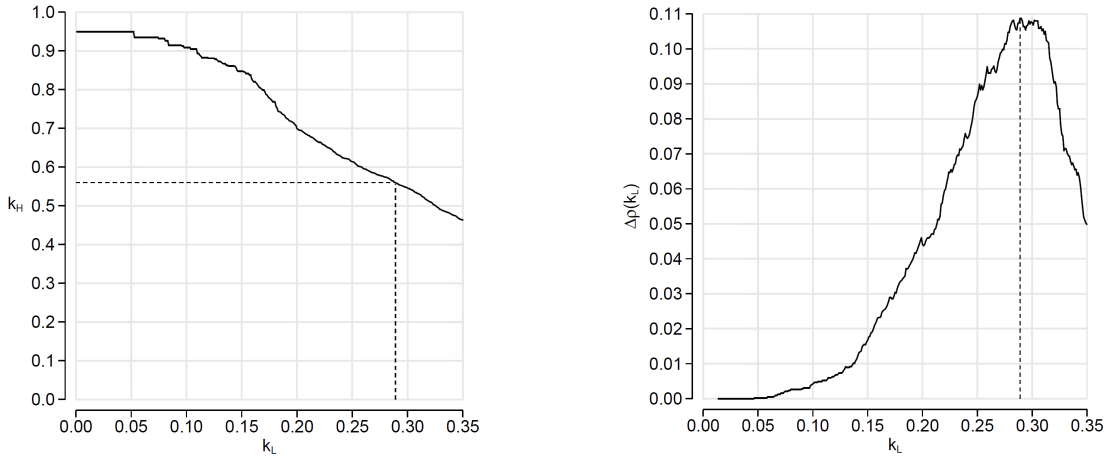


Figure 9: Thresholds k_L and k_H (left) minimizing antibiotic use for fixed number of correctly treated bacterial UTIs and optimal reduction in prescribing as function of k_L (right) with redistribution from pregnant women prohibited.

If physicians immediately prescribe to pregnant women, not because they fail to properly evaluate risk, but rather because they assign a high sickness cost, it is then indeed optimal to prescribe even at low risk of bacterial infection. We can identify patient pregnancy in our data by claims for periodic examinations during pregnancy required by law. Pregnant women account for 28 percent of patients in the hold-out sample.[13] To examine if our policy results are driven by redistributing

[13]Pregnant women have higher risk of UTI and guidelines recommend that physicians always acquire urine culture from pregnant women showing UTI symptoms. Combined with a higher sickness cost, it is not surprising that they

away from low risk but high cost pregnant women, we implement a modification to the prescription rule ρ^S in that no prescription can be removed from a pregnant woman even if predicted risk is below the no-prescription threshold, $m(x_i) < k_L$.

Figures 9 and 10 show redistribution based on the updated switching rule. As can be seen, our main results remain intact, in fact slightly increasing the potential for improvement. This finding suggests that physicians do not base their prescription decisions on higher sickness cost for pregnant women but their evaluation of risk. Therefore, we conclude that our welfare improvement findings hold and are not diminished by redistributing prescriptions away from high sickness cost patients.

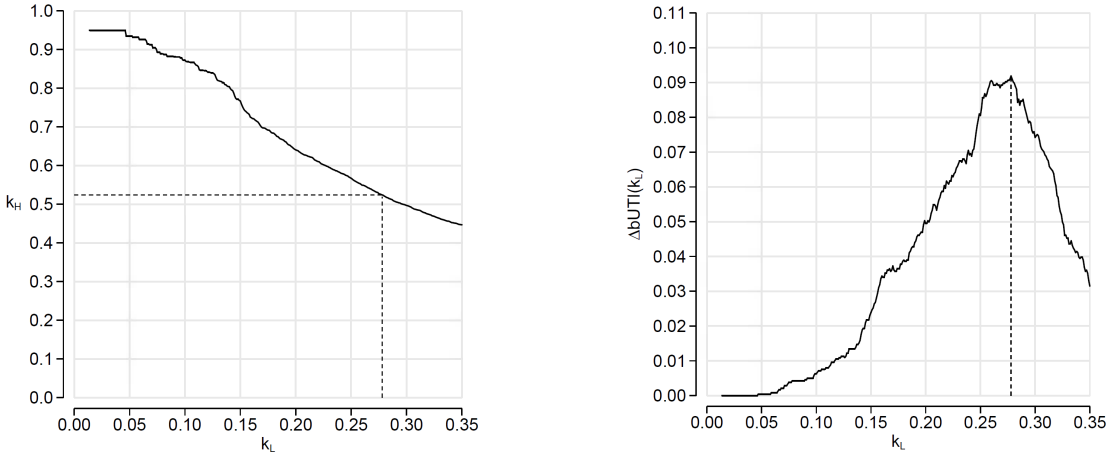


Figure 10: Thresholds k_L and k_H (left) maximizing treated bacterial UTI for fixed total amount of prescriptions and increase in treated bacterial UTI as function of k_L (right) with redistribution from pregnant women prohibited.

7.2 How Much Data are Needed?

We have seen that machine prediction can achieve significant improvements in antibiotic prescribing. However, the richness and interconnectedness of our data is unique to the Danish context. Some other Scandinavian countries collect and make similar linkable register data available but it is an important question how easily our analysis can be ported to the majority of countries in which data availability is inferior due to restrictive data protection laws or lack of data collection. What are the data requirements to attain comparable machine predictions and improvements in prescribing? We consider this issue by using two reduced sets of variables in X to rerun our analysis based on the exact same training and hold out observations.

are overrepresented in our data (Foxman, 2002).

For the first reduced set, we keep the subset of patient characteristics: gender, civil status, family type, income, and age, and combine them with patients' prescription histories, their laboratory test histories, and past hospitalizations. These data would typically be available in national health insurance systems or health insurance companies in developed countries. The resulting AUC for this set X is 0.7273, comparable to prediction quality based on the full set of explanatory variables. Applying the analogous redistribution mechanisms, we can achieve a maximum improvements a 10.1 percent reduction in prescribing, holding the number of correctly treated bacterial UTI constant, and a 7.9 increase in correctly treated bacterial UTI, holding the total number of prescriptions constant. These results are not distinguishable from our main results and, hence, our approach to improve antibiotic prescribing shows promise for many other developed healthcare systems.

The second set of explanatory variables contains data on the same patient characteristics as above, which are typically broadly available, and patients' prescription histories, arguably the most easily tractable medical histories for individual patients. Again, we achieve a comparable prediction quality based on the standard machine learning measure, an AUC equal to 0.7242. However, applying our redistribution rules, we find that the improvements we can achieve have decreased. The maximum improvements are now a 7.7 percent reduction in prescribing, holding the number of correctly treated bacterial UTI constant, and a 5.9 increase in correctly treated bacterial UTI, holding the total number of prescriptions constant. These are still improvements so that considering using machine prediction to improve antibiotic prescribing may be useful in a large number of other institutional circumstances.

Analyzing improvements based on these alternative predictor subsets illustrates an important point. While overall prediction quality, measured by the AUC, remains unchanged, the comparison with human decisions does not. This shows that implementing decisions based solely on machine prediction is problematic as the scope for improvements crucially depends on the quality of human experts' risk assessment over the full support of risks as well as on their payoff functions.

8 Conclusion

In this paper we have shown that machine prediction can reduce antibiotic prescribing in economically significant ways while maintaining treatment quality. Antibiotic prescribing under uncertainty about the cause of infection is a high impact policy problem given the empirical relevance of increasing antibiotic resistance, which is partly due to wasteful antibiotic prescribing. While we show that

implementing machine predictions leads to improvements over physicians' prescription decisions, machine predictions alone cannot achieve improvements. Physician expertise, that is their observation and assessment of factors that are unobservable to the machine learning algorithm, remains crucial in prescription choices. Still, for a large interval of machine-predicted risks, using machine prediction leads to unambiguous welfare improvement.

One promising avenue for further research is the combination of machine predictions with results from in-clinic, imperfect diagnostics for bacterial infection causes, such as nitrate stick and microscopy. We also omit an important dimension of antibiotic prescribing, the choice of molecule. It is an interesting question beyond the scope of this paper, to what extent machine prediction of bacterial species and molecule-specific resistances are able to further improve prescription choices. This is left for future research. Finally, our machine prediction results could be used to assist physicians in their decision-making, for example by providing physicians with a machine predicted risk at every prescription occasion. A full assessment of the equilibrium effects of such an assistance will require interventions in the field combined with machine prediction, a promising avenue for future research to further improve antibiotic prescribing.

One limitation is that we consider only first consultation prescription occasions in which a laboratory test was ordered. Yet, given the problem that UTI typically must be treated quickly and laboratory testing takes considerable time our analysis still holds empirical relevance. In Ribers and Ullrich (2018) we consider a more general setting capturing all observed prescription choice occasions throughout patients' treatment spells in a structural dynamic model that endogenizes both the prescription and test decision. Extending the machine prediction approach considered here to a similarly general setting is beyond the scope of this paper. One promising avenue to tackle this issue in practical implementation is randomization of additional laboratory testing. This being done, machine prediction could be evaluated against physician decisions on a generalizable sample of the population.

Our analysis shows the potential of the careful use of machine learning methods in specific healthcare treatment choice situations. The quality of prediction algorithms and data availability are improving at a rapid pace. Yet, the challenge of evaluating machine prediction against human decisions in the presence of unobservables and decision-makers' heterogenous objective functions will remain and needs to be solved for each implementation setting.

A Training and Hold Out Partitions

Table 2 Bacterial Isolates in the Training and Hold Out Partitions

	All		Training Data		Hold Out	
Genus	Freq	Pct.‡	Freq	Pct.‡	Freq	Pct.‡
–	75,360	–	63,717	–	11,643	–
Escherichia	20,990	64.2	17,462	64.1	3,528	64.5
Enterococcus	3,000	9.2	2,462	9.0	538	9.8
Staphylococcus	2,514	7.7	2,227	8.2	287	5.3
Streptococcus	1,857	5.7	1,600	5.9	257	4.7
Klebsiella	1,793	5.5	1,412	5.2	381	7.0
Proteus	707	2.2	567	2.1	140	2.6
Citrobacter	499	1.5	394	1.5	105	1.9
Enterobacter	403	1.2	323	1.2	80	1.5
Pseudomonas	382	1.2	312	1.2	70	1.3
Other	553	1.7	470	1.7	83	1.5
Observations	108,058	32,698	90,946	27,229	17,112	5,469

‡ Percentages are reported as shares of isolated bacteria.

B Machine Learning Performance

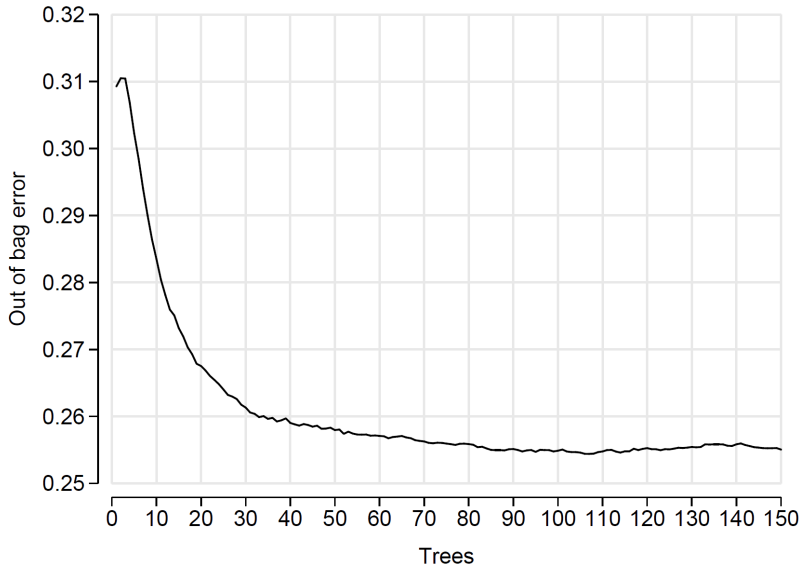


Figure 11: The necessary number of trees can be determined by the mean prediction error on a separate cross-validation dataset split from the training data. However, since random forests bootstrap the training data when constructing each individual tree, one can evaluate an equivalent prediction error at an observation by computing the mean prediction error using only trees in the forest that did not include the observation in question. This is the out of bag error (OOB error) which has been shown to give almost identical results to cross-validation without the inefficiency of splitting the data (Hastie et al. 2009). The figure above shows the evolution of the OOB error for our algorithm as the number of trees increase. The OOB error stabilizes at around 100 trees, which is what we use in our final specification of the random forest algorithm.

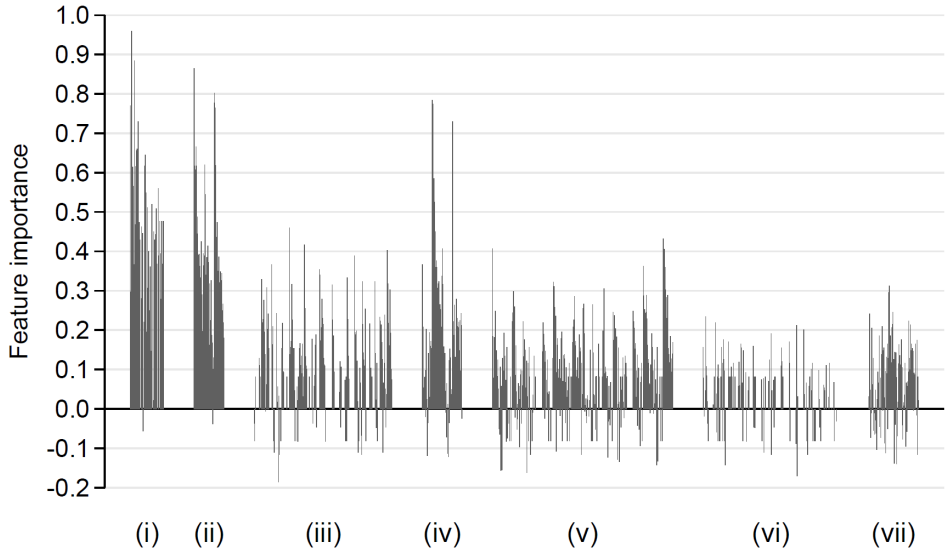


Figure 12: It is instructive to see which variables most influence prediction as shown by the feature importance plot above. The feature importance is computed for each variable independently as the increase in the OOB error when the values of a variable are permuted for the out of bag observations. Variables that are 0 or negative should be considered to have no impact on prediction as random permutation did not decrease prediction. We have too many variables to list feature importance and have instead marked them in groups containing (i) patient characteristics and timing, (ii) patient past prescriptions, (iii) patient past resistance test results, (iv) patient past hospitalization, (v) household members' past prescriptions, (vi) household members' past resistance test results, and (vii) household members' past hospitalizations. The most notable result is that patient past test results do not appear to be important features. This, however, is most likely because the past test results are highly correlated with past consumption. Correlated features reduce each others importance in feature importance plots and we cannot conclude that past test results are less important based on a feature importance plot alone.

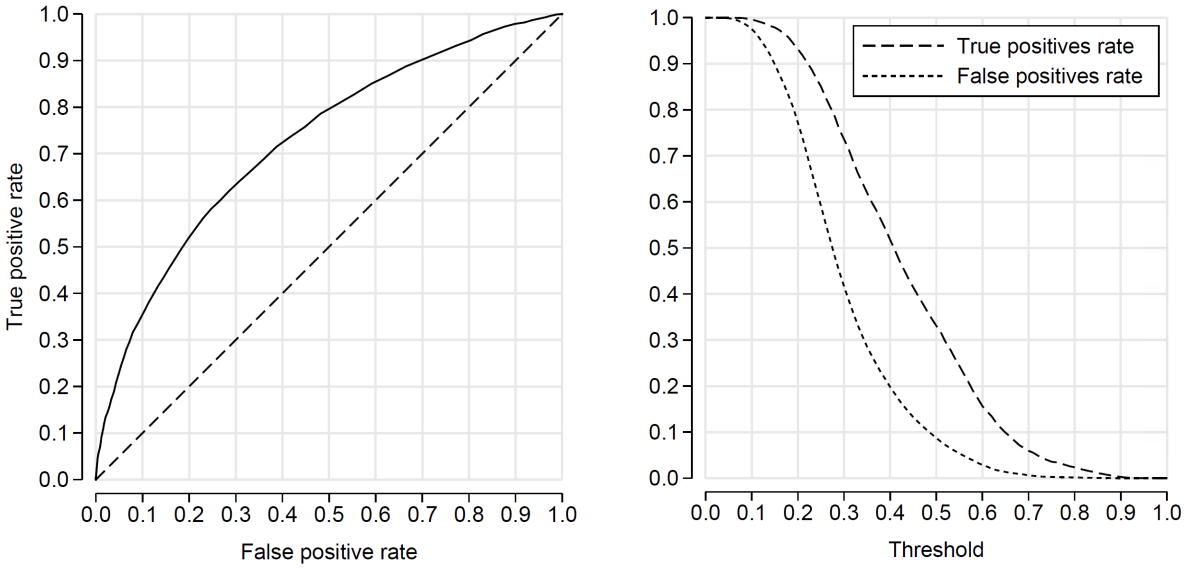


Figure 13: The receiver operating curve (left) shows the model's achievable true positive rate versus the tradeoff in the false positive rate (right) as the classification thresholds, the predicted risk at which binary bacterial outcome is classified, varies over $[0, 1]$. The 45 degree line represents random guessing while points above the diagonal represent better than random prediction. The area under the curve, a common measure of precision, computes to 0.726, comparable to Kleinberg et al. (2017), who report an AUC of 0.707. Similar to the ROC, the interpretation of the AUC is that 0.5 represents random guessing while values closer to 1 represents better prediction.

References

[1] Albert, Jason (2015), "Strategic dynamics of antibiotic use and the evolution of antibiotic-resistant infections," mimeo.

[2] Arnold, Sandra R. and Sharon E. Straus (2005), "Interventions to improve antibiotic prescribing practices in ambulatory care," *The Cochrane Library.*

[3] Athey, Susan (2018), "The impact of machine learning on economics," in *The economics of artificial intelligence: an agenda* ed. Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb, University of Chicago Press.

[4] Bennett, Daniel, Hung, Che-Lun, and Tsai-Ling Lauderdale (2015), "Health care competition and antibiotic use in Taiwan," *The Journal of Industrial Economics*, Vol. 63, No. 2, pp. 371-393.

[5] Bjerrum, Lars and Morten Lindbæk (2015), "Which treatment strategy for women with symptoms of urinary tract infection?," *BMJ*, Vol. 351, pp. 1-2.

[6] Breiman, Leo (2001), "Random forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

[7] Brown, Gardner and David F. Layton (1996), "Resistance economics: social cost and the evolution of antibiotic resistance," *Environment and Development Economics*, Vol. 1, No. 3, pp. 349-355.

[8] Butler, Christopher C., Simpson, Sharon A., Dunstan, Frank, Rollnick, Stephen, Cohen, David, Gillespie, David, Evans, Meirion R., Alam, M. Fasihul, Bekkers, Marie-Jet, Evans, John, Moore, Laurence, Howe, Robin, Hayes, Jamie, Hare, Monika, and Kerenza Hood (2012), "Effectiveness of multifaceted educational programme to reduce antibiotic dispensing in primary care: practice based randomised controlled trial," *BMJ*, Vol. 344.

[9] Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016), "Productivity and selection of human capital with machine learning," *American Economic Review*, Vol. 106, No. 5, pp. 124-127.

[10] Chen, Jonathan H. and Steven M. Asch (2017), "Machine learning and prediction in medicine-beyond the peak of inflated expectations," *New England Journal of Medicine*, Vol. 376, No. 26, pp. 2507-2509.

[11] Currie, Janet, Wanchuan Lin, and Juanjuan Meng (2014), "Addressing antibiotic abuse in China: an experimental audit study," *Journal of Development Economics*, Vol. 110, pp. 39-51.

[12] Currie, Janet and W. Bentley MacLeod (2017), "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians," *Journal of Labor Economics*, Vol. 35, No. 1, pp. 1-43.

[13] Danish Ministry of Health (2017), National handlingsplan for antibiotika til mennesker. Tre målbare mål for en reduktion af antibiotikaforbruget frem mod 2020.

[14] Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan (2016), "Quality and accountability in health care delivery: audit-study evidence from primary care in India," *American Economic Review*, Vol. 106, No. 12, pp. 3765-3799.

[15] Devillé, Walter L.J.M., Joris C. Yzermans, Nico P. van Duijn, P. Dick Bezemer, Daniëlle A.W.M. van der Windt, and Lex M. Bouter (2004), "The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy," *BMC Urology*, Vol. 4, No. 4, pp. 1-14.

[16] Elbasha, Elamin H. (2003), "Deadweight loss of bacterial resistance due to overtreatment," *Health Economics*, Vol. 12, No. 2, pp. 125-138.

[17] Eswaran, Mukesh and Nancy Gallini (2018), Can competition extend the golden age of antibiotics?, mimeo.

[18] Flores-Mireles, Ana L., Jennifer N. Walker, Michael Caparon, and Scott J. Hultgren (2015), "Urinary tract infections: epidemiology, mechanisms of infection and treatment options," *Nature Reviews Microbiology*, Vol. 13, pp. 269-284.

[19] Foxman, Betsy (2002), "Epidemiology of urinary tract infections: incidence, morbidity, and economic costs," *The American Journal of Medicine*, Vol. 113, No. 1, Suppl. 1, pp. 5-13.

[20] Goossens, Herman, Matus Ferech, Robert Vander Stichele, and Monique Elseviers (2005), "Outpatient antibiotic use in Europe and association with resistance: a cross-national database study", *The Lancet*, 365(9459), 579-587.

[21] Grigoryan, Larissa, Trautner, Barbara W., and Kalpana Gupta (2014), "Diagnosis and management of urinary tract infections in the outpatient setting: a review," *JAMA*, Vol. 312, No. 16, pp. 1677-1684.

[22] Hallsworth, Michael, Tim Chadborn, Anna Sallis, Michael Sanders, Daniel Berry, Felix Greaves, Lara Clements, and Sally C. Davies (2016), "Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial," *The Lancet*, Vol. 387, pp. 1743-1752.

[23] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The elements of statistical learning: data mining, inference, and prediction*, 2nd Edition, New York: Springer.

[24] Herrmann, Markus and Gérard Gaudet (2009), "The economic dynamics of antibiotic efficacy under open access," *Journal of Environmental Economics and Management*, Vol. 57, No. 3, pp. 334-350.

[25] Hooton, Thomas M. (2012), "Uncomplicated urinary tract infection," *New England Journal of Medicine*, Vol. 366, No. 11, pp. 1028-1037.

[26] Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction policy problems," *American Economic Review*, Vol. 105, No. 5, pp. 491-495.

[27] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017), "Human decisions and machine predictions," *Quarterly Journal of Economics*, Vol. 133, No. 1, pp. 237-293.

[28] Kwon, Illoong and Daesung Jun (2015), "Information disclosure and peer effects in the use of antibiotics," *Journal of Health Economics*, Vol. 42, pp. 1-16.

[29] Laxminarayan, Ramanan and Gardner M. Brown (2001), "Economics of antibiotic resistance: a theory of optimal use," *Journal of Environmental Economics and Management*, Vol. 42, No. 2, pp. 183-206.

[30] Laxminarayan, Ramanan, Adriano Duse, Chand Wattal, Anita K.M. Zaidi, Heiman F.L. Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M. Gould, Herman Goossens, Christina Greko, Anthony D. So, Maryam Bigdeli, Göran Tomson, Will Woodhouse, Eva Ombaka, Arturo Quizhpe Peralta, Farah Naz Qamar, Fatima Mir, Sam Kariuki, Zulfiqar A. Bhutta, Anthony Coates, Richard Bergstrom, Gerard D. Wright, Eric D. Brown, and Otto Cars (2013), "Antibiotic resistance – the need for global solutions," *The Lancet Infectious Diseases Commission*, pp. 1-42.

[31] Laxminarayan, Ramanan and Martin L. Weitzman (2002), "On the implications of endogenous resistance to medications," *Journal of Health Economics*, Vol. 21, No. 4, pp. 709-718.

[32] Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014), "The parable of google flu: traps in big data analysis," *Science*, Vol. 343, No. 6176, pp. 1203-1205.

[33] Obermeyer, Ziad and Ezekiel J. Emanuel (2016), "Predicting the future – big data, machine learning, and clinical medicine," *New England Journal of Medicine*, Vol. 375, No. 13, pp. 1216-1219.

[34] Ribers, Michael and Hannes Ullrich (2018), "Prescribing antibiotics under uncertainty about resistance," mimeo.

[35] Rudholm, Niklas (2002), "Economic implications of antibiotic resistance in a global economy," *Journal of Health Economics*, Vol. 21, No. 6, pp. 1071-1083.

[36] Schieve, Laura A., Arden Handler, Ronald Hershow, Victoria Persky, and Faith Davis (1994), "Urinary tract infection during pregnancy: its association with maternal morbidity and perinatal outcome," *American Journal of Public Health*, Vol. 84, No. 3, pp. 405-410.

[37] Schwandt, Hannes (2018), "The lasting legacy of seasonal influenza: in-utero exposure and labor market outcomes," CEPR Discussion Paper No. 12563.

[38] World Health Organization (2012), The evolving threat of antimicrobial resistance: options for action, Geneva, Switzerland.

[39] World Health Organization (2014), Antimicrobial Resistance: Global Report on Surveillance, Geneva, Switzerland.

[40] Yip, Winnie Chi-Man, William Hsiao, Qingyue Meng, Wen Chen, and Xiaoming Sun (2010), "Realignment of incentives for health-care providers in China," *The Lancet*, Vol. 375, pp. 1120-1130.